A COMPARISON BETWEEN THE VARYING-THRESHOLDS MODEL AND QUANTILE REGRESSION

Niccolò Ducci [1], Leonardo Grilli[2] and Marta Pittavino[2]

[1] Agenzia delle Entrate (e-mail: niccolo.ducci.it@gmail.com)

[2] Department of Statistics, Computer Science, Applications ”Giuseppe Parenti”, University of Florence (e-mail: leonardo.grilli@unifi.it, marta.pittavino@unifi.it)

ABSTRACT: The varying-thresholds model is a new modelling approach capable of estimating the whole conditional distribution of a response variable in a regression setting. The varying-thresholds model can be used for continuous, ordinal and count responses. Conditional quantiles estimated through the varying-thresholds method are compared to those of quantile regression. The comparison is based on models' simulations to assess the performance of the two methodologies regarding the coverage and width of prediction intervals. The simulation study encompasses eight different settings with several functional forms and types of errors. In addition, a discrete variation of the continuous ranked probability score is proposed as a way to choose the best link function for the binary models used to estimate the varying-thresholds model. The comparison shows that the varying thresholds model performs better whenever the functional form of the true data generating model is non-linear.

KEYWORDS: varying-thresholds model, quantile regression, robit, prediction intervals, continuous ranked probability score

1 The Varying-Thresholds Model

The varying-thresholds model is a novel methodology proposed by Tutz, 2021 that can estimate the whole conditional distribution of a response variable in a regression setting. Estimating the conditional distribution allows one to obtain values of interest such as the conditional expected value, standard error, or quantiles. The general form of the Varying-Thresholds Model can be written as follows:

$$P(Y > \theta \,|\, \mathbf{x}) = F(\,\eta(\theta, \mathbf{x})\,) \tag{1}$$

where Y is the response variable, $\mathbf{x}$ is a vector of covariates, F is a distribution function and $\eta(\theta, \mathbf{x})$ is a predictor function. The predictor function can take

Table 1. *All types of data generating models used in the comparison between quantile regression and the varying-thresholds model. Every model comprise a single covariate and a response variable.*

Model	*Functional Form*	Error Distribution	Covariate Distribution
Model 1	$\beta_0 + \beta_1 x$	$\varepsilon_N \sim N(0,1)$	$X \sim N(5,1)$
Model 2	$\beta_0 + \beta_1 x$	$\varepsilon_{\chi^2} \sim \chi^2(df = 3)$	$X \sim N(5,1)$
Model 3	$\beta_0 + \beta_1 x$	$\varepsilon_N \sim e^{(x-5)} \cdot N(0,1)$	$X \sim N(5,1)$
Model 4	$\beta_0 + \beta_1 x$	$\varepsilon_N \sim e^{(5-x)} \cdot N(0,1)$	$X \sim N(5,1)$
Model 5	$\beta_0 + \beta_1 x + \beta_2 x^2$	$\varepsilon_N \sim N(0,1)$	$X \sim U(-2,12)$
Model 6	$\beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$	$\varepsilon_N \sim N(0,1)$	$X \sim U(-3,8)$
Model 7	$\beta_0 + \beta_1 x$	$\varepsilon_{CN} \sim 0.9N(0,5)+$ $0.1N(50,5)$	$X \sim N(0,5)$
Model 8	$\beta_0 + \beta_1 x + \beta_2 x^2$	$\varepsilon_t \sim t(df = 3)$	$X \sim U(0,6)$

many forms: linear, non-linear or non-parametric. In this work, we consider a single covariate x and adopt a linear specification $\eta(\theta, \mathbf{x}) = \beta_0^\theta + \beta_1^\theta x$. The response variable Y can be ordinal or continuous. The varying-thresholds model is estimated using a series of binary regression models: for every threshold θ in a prespecified grid of values, the response variable Y is dichotomized to become binary, then a model is fitted to the data as described in equation 1. This method allows for the estimation of varying coefficients, indexed by θ, that are then used to compute the conditional distribution of the response variable [*].

2 Data Generating Models and Simulation Settings

The varying-thresholds model and quantile regression are compared using a variety of error assumptions and different functional forms. Quantile regression is fitted as $Q_{Y|x}(\theta) = \beta_0(\theta) + \beta_1(\theta)x$, likewise the varying-thresholds model is estimated using the predictor function $\eta(\theta, \mathbf{x}) = \beta_0^\theta + \beta_1^\theta x$. All the data generating models are reported in Table 1. Model's errors mimic the

[*] Note that, even if the predictor is linear, the binary response model is repeatedly fitted with different thresholds, thus the regression function is estimated in a data-driven way.

latent response approach , i.e. $Y^* = functional form + error$ and $Y = 1$ if and only if $Y^* > 0$, e.g., a model with normally distributed error corresponds to the probit model. The errors are always standardized to ensure comparability of the regression coefficients. Quantile regression and the varying-thresholds model are compared through the empirical coverage of their estimated prediction intervals computed at a $(1-\alpha) = 80\%$ level conditioned on a given value of $X = x$. This interval is computed by estimating the first and ninth conditional decile. The empirical coverage is calculated through a simulation. The simulation has 1000 iterations, each time a different sample of $n = 1000$ observations is drawn from the generating model. After each iteration the two methodologies compute the intervals; then, a new observation is sampled from the generating model; the proportion of times the new observation falls within the prediction interval is the empirical coverage level. Quantile regression is estimated with the R package `quantreg`, Koenker, 2022.

3 Simulation Results and Link Selection

Table 2 reports the results of the simulations for prediction intervals conditioned on the median value of X. The comparison shows that the varying-thresholds model performs better whenever the functional form of the true data generating model is non-linear. The lack of assumptions about the functional relationship makes the varying-thresholds model a very flexible approach, capable of detecting non-linear effects without specifying a non-linear effect in the predictor function $\eta(\theta, \mathbf{x})$. If the functional relationship between variables is known in advance and it is correctly specified quantile regression generally yields better results. The choice of the link function for the binary models used to estimate the varying-thresholds model is crucial; a discrete approximation of the continuous ranked probability score (CRPS), Jordan *et al.*, 2019; Gneiting & Raftery, 2007, is used to select the best link function. Both out-of-sample or in-sample approaches seems to be valid with this metric. In $Model 8$ the robit link function with three degrees of freedom is selected through the CRPS and yields better results than other links.

4 Conclusions

The varying-thresholds model performs better, regarding prediction intervals, than quantile regression when there are non-linear effects and the relationship between variables is not correctly specified. Link function selection for the binary models' estimation method can be facilitated using the CRPS. Areas of

Table 2. *Empirical coverage and average width of prediction intervals at* 80% *level on* 1000 *simulations from* $Model 1-8$ *at the median value of* X. *The varying-thresholds model is fitted with probit link function except for* $Model 8$ *where it is fitted with robit*[a] *link function with three degrees of freedom.*

Model	*Quantile Regression*		*Varying-Thresholds Model*	
	Coverage	*Avg. Width*	*Coverage*	*Avg. Width*
Model 1	0.783	2.562	0.783	2.567
Model 2	0.820	5.670	0.822	5.788
Model 3	0.926	3.759	0.930	4.124
Model 4	0.937	3.755	0.939	4.121
Model 5	0.704	4.651	0.865	3.306
Model 6	1.000	8.559	0.883	3.354
Model 7	0.801	33.221	0.844	37.694
Model 8	0.649	7.964	0.810	3.442

[a] The robit link function is related to the t-distribution, see Liu, 2004.

future research may include different types of response variables such as count and ordinal data.

References

Gneiting, T., & Raftery, A. E. 2007. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, **102**(477), 359–378.

Jordan, A., Krüger, F., & Lerch, S. 2019. Evaluating Probabilistic Forecasts with scoringRules. *Journal of Statistical Software*, **90**(12), 1–37.

Koenker, R.W. 2022. *quantreg: Quantile Regression*. R package version 5.94.

Liu, C. 2004. *Robit Regression: A Simple Robust Alternative to Logistic and Probit Regression*. John Wiley & Sons, Ltd. Chap. 21, pages 227–238.

Tutz, G. 2021. *Flexible Predictive Distributions from Varying-Thresholds Modelling*. arXiv:2103.13324.